Project 2: Iowa Housing

By Tanupong Rattanasawatesun

Overview

Overview: Aimes Iowa housing

Aimes Iowa Housing datasets contain 2051 row of datapoints. While each row represents a house sold from 2006 to 2010, each columns contain features or characteristics of that house.

Linear Regression model is used to predict the sale price from those features. The model is optimized by feature engineering and subset feature selection. The model is evaluated and analyzed to give the factor that affect the housing price both positively and negatively.

This model can be utilized in many ways such as to predict the price of the house or to find a best way to spend money in house investment.

Problem statement:

For customers who want to sell their house, what is an estimated current sale price ?

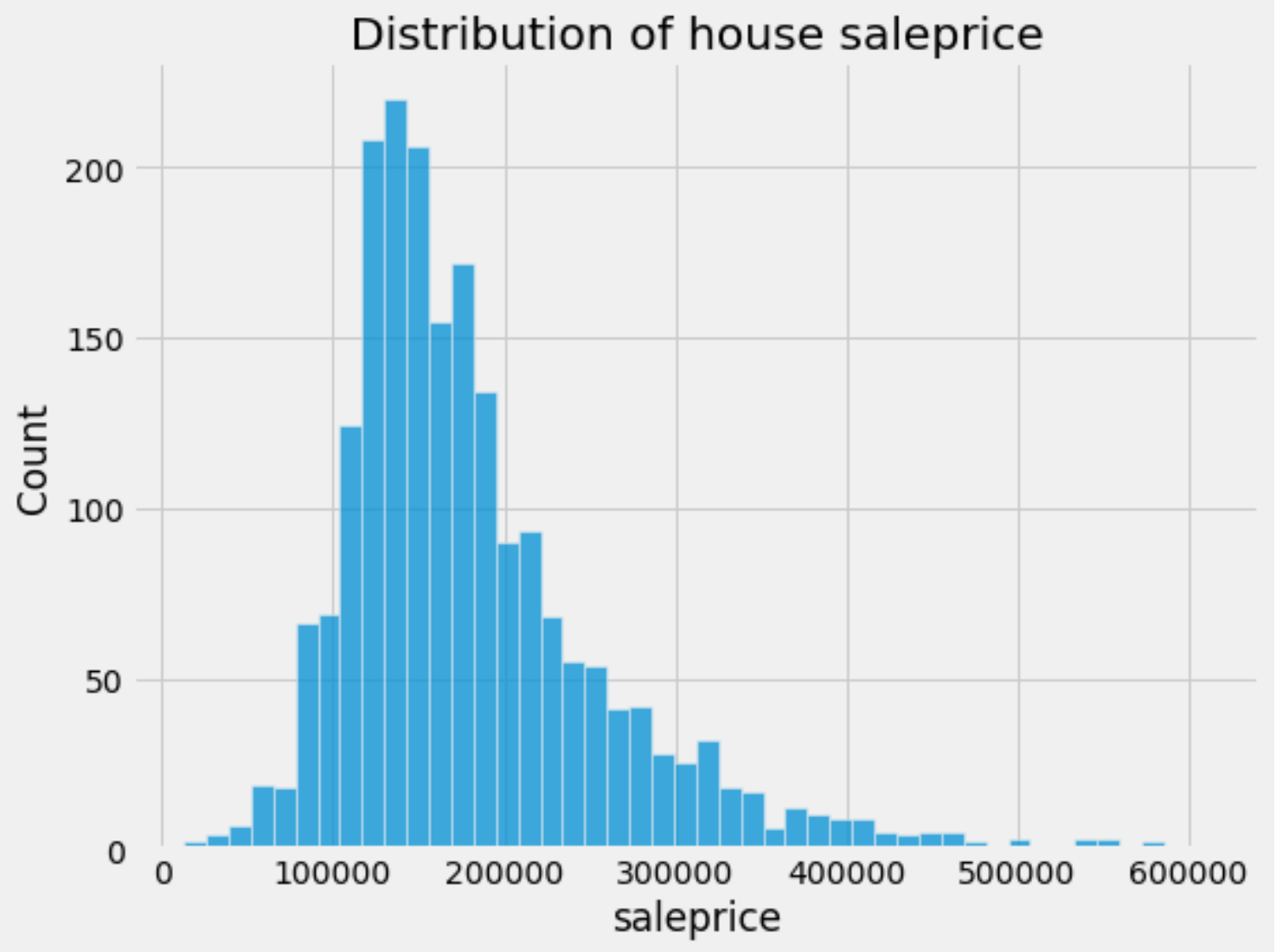
And what should be any improvements that can raise the price up?

Exploratory Data Analysis

Distribution of Sale Price

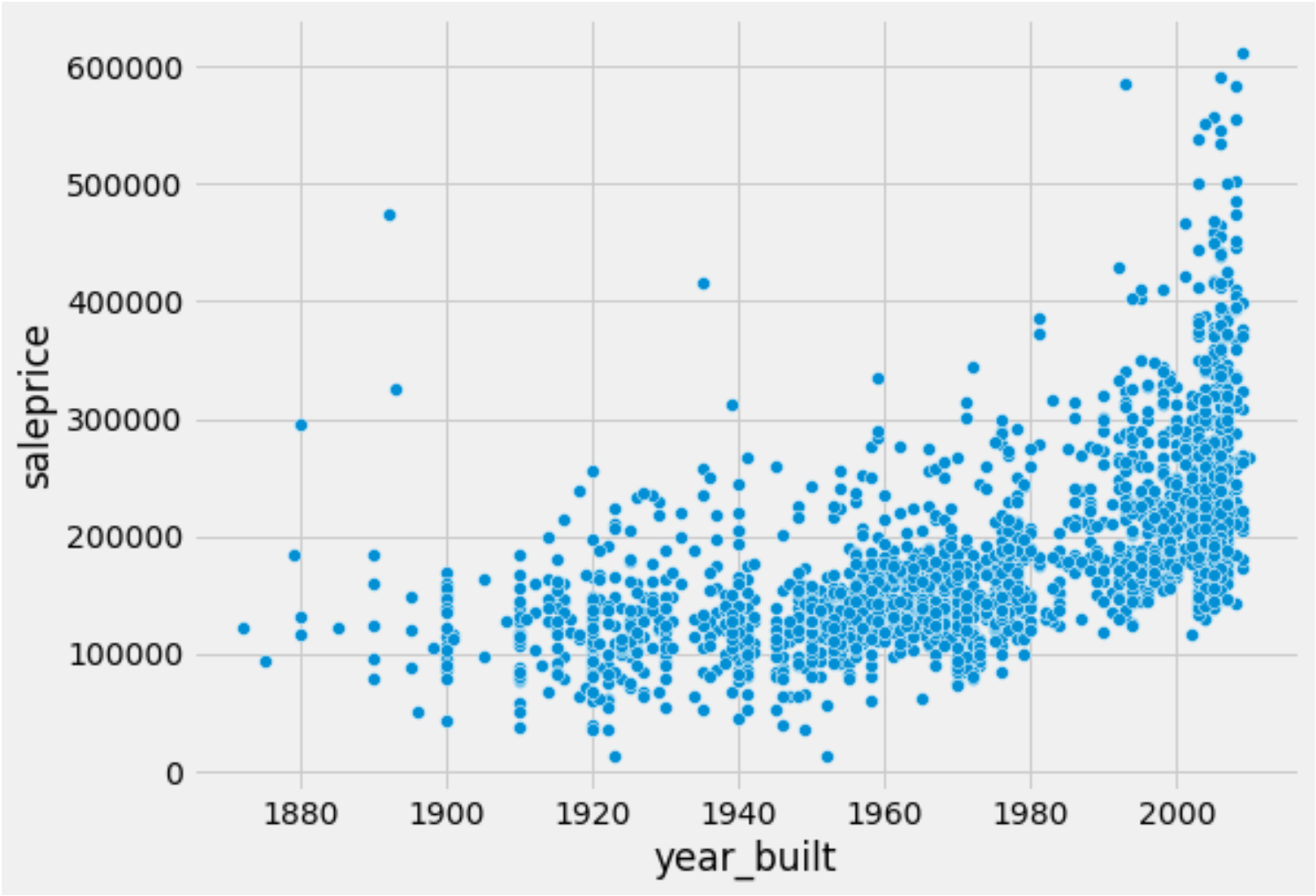
The range of sale price is from around 10000 USD to 600000 USD.

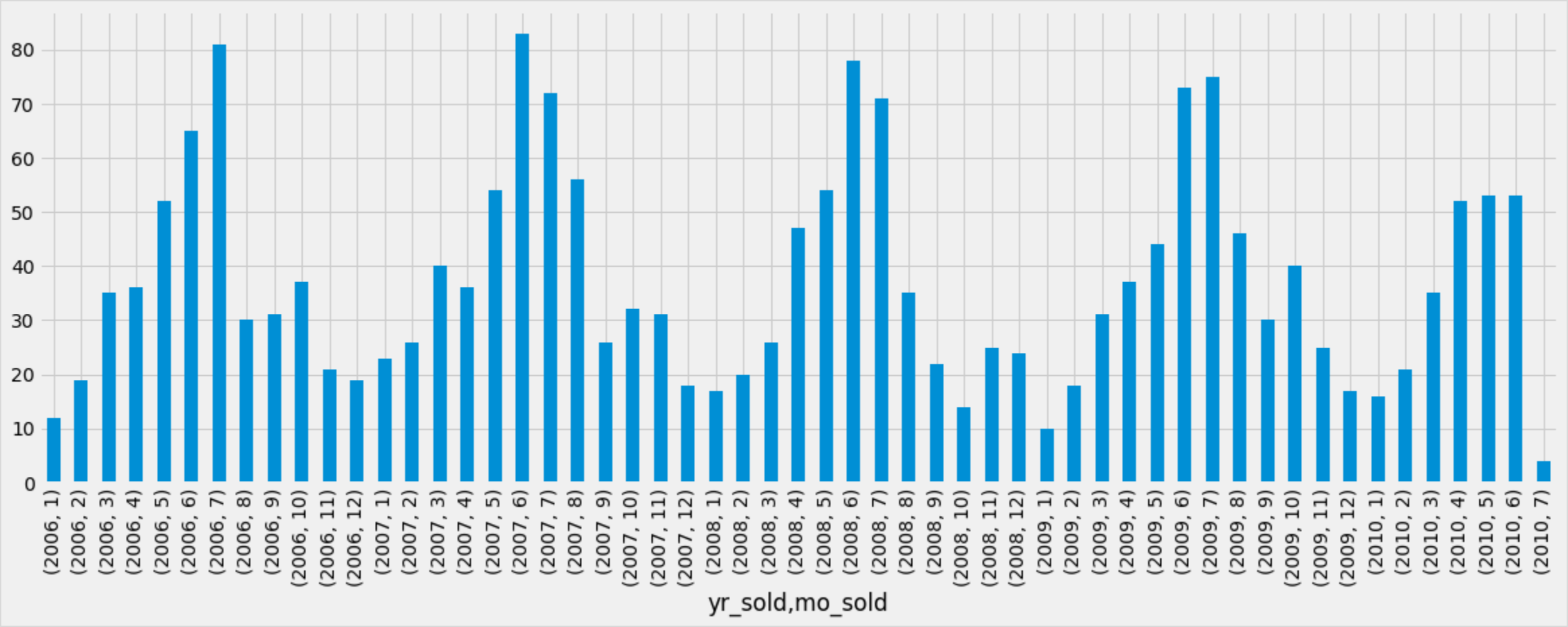
The average sale price is about 181000 USD.



SalePrice vs Year built

The house built before year 2000 usually has the price in between 100000 – 300000 USD



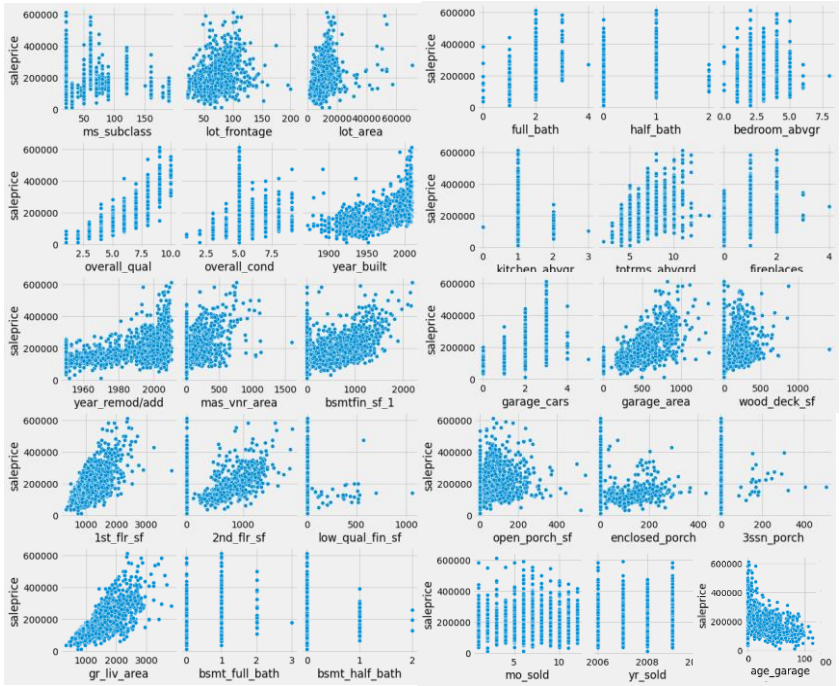


Time of the house-selling in 2006-2010

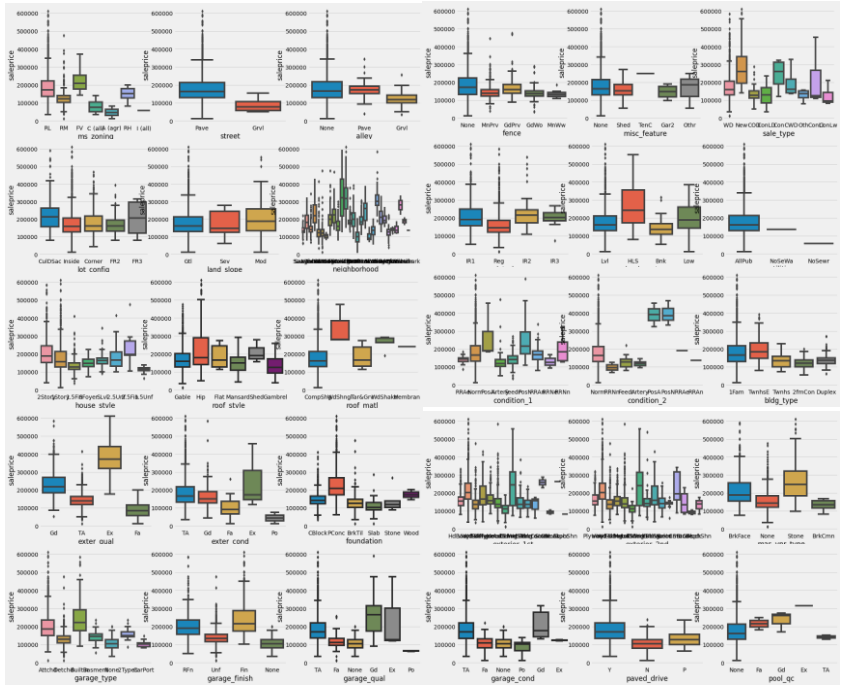
There is a peak in around May-August of each year. People tend to buy/sell houses in the middle of the year

Data Modeling

How to choose features



Numerical Features



Categorical Features

Features selection

Numerical Features
Total Area in Sq.Ft.
Year Built
Overall Qualty
Basement finished area in Sq.Ft.
Above grade (ground) living area in Sq.Ft.
Total Floor Area in Sq.Ft.
Masonry veneer area in Sq.Ft.
Overall Condition
Garage Area in Sq.Ft.
Lot Area in Sq.Ft.
Exterior Quality
Kitchen Quality
has garage

Categorical features
Neighborhood
MS Zoning
MS Subclass
Exterior covering on house
Masonry veneer type
Roof Style
Roof material
Bldg Type
Heating
Basement Exposure
Rating of basement finished area
Garage Finish
Home Functionality
Flatness of the property
Lot configuration
Proximity to various conditions

Result & Conclusion

Model Evaluation

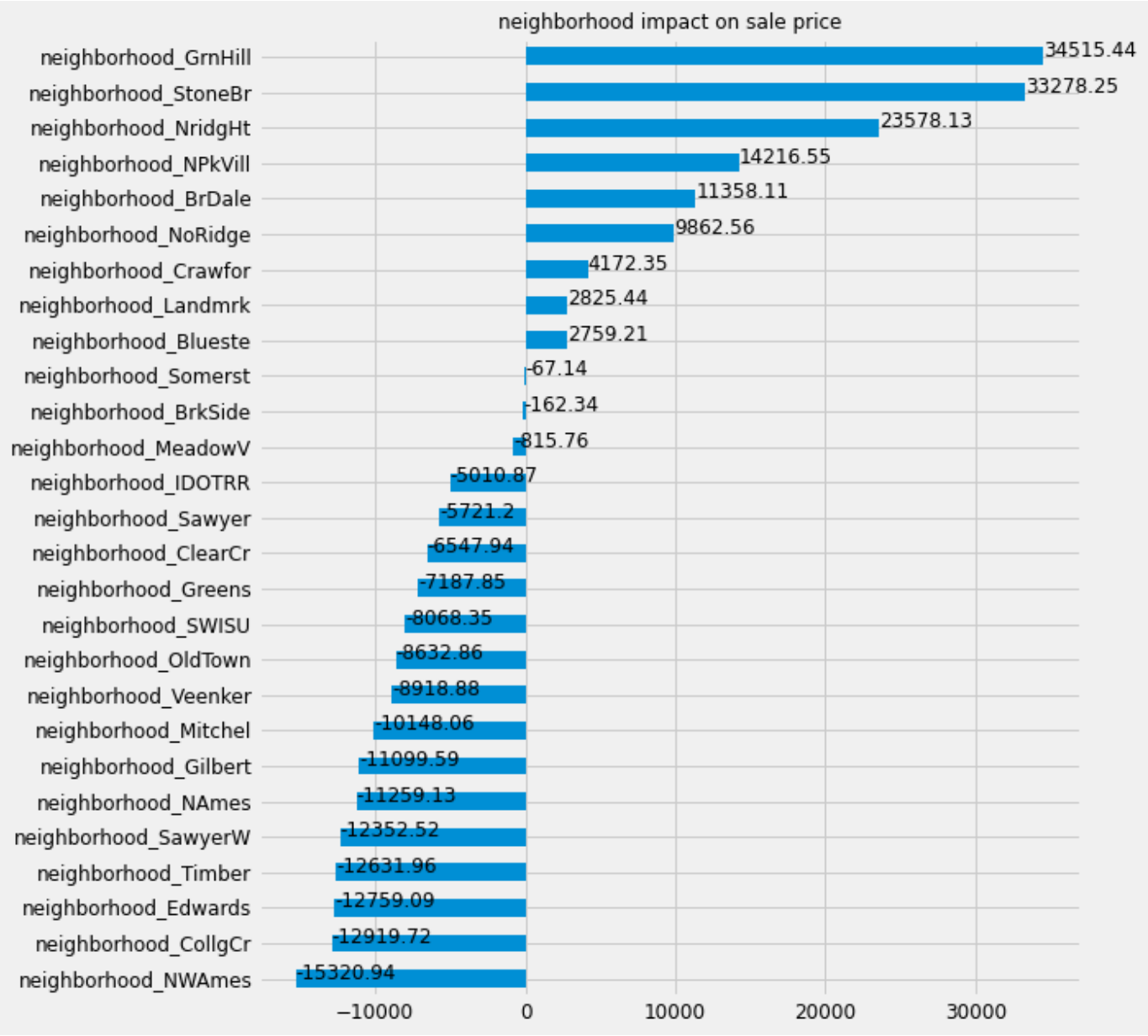
Model Performance

Training data
R2 Score: 0.92244
Root Mean Square Error: 22075.1305

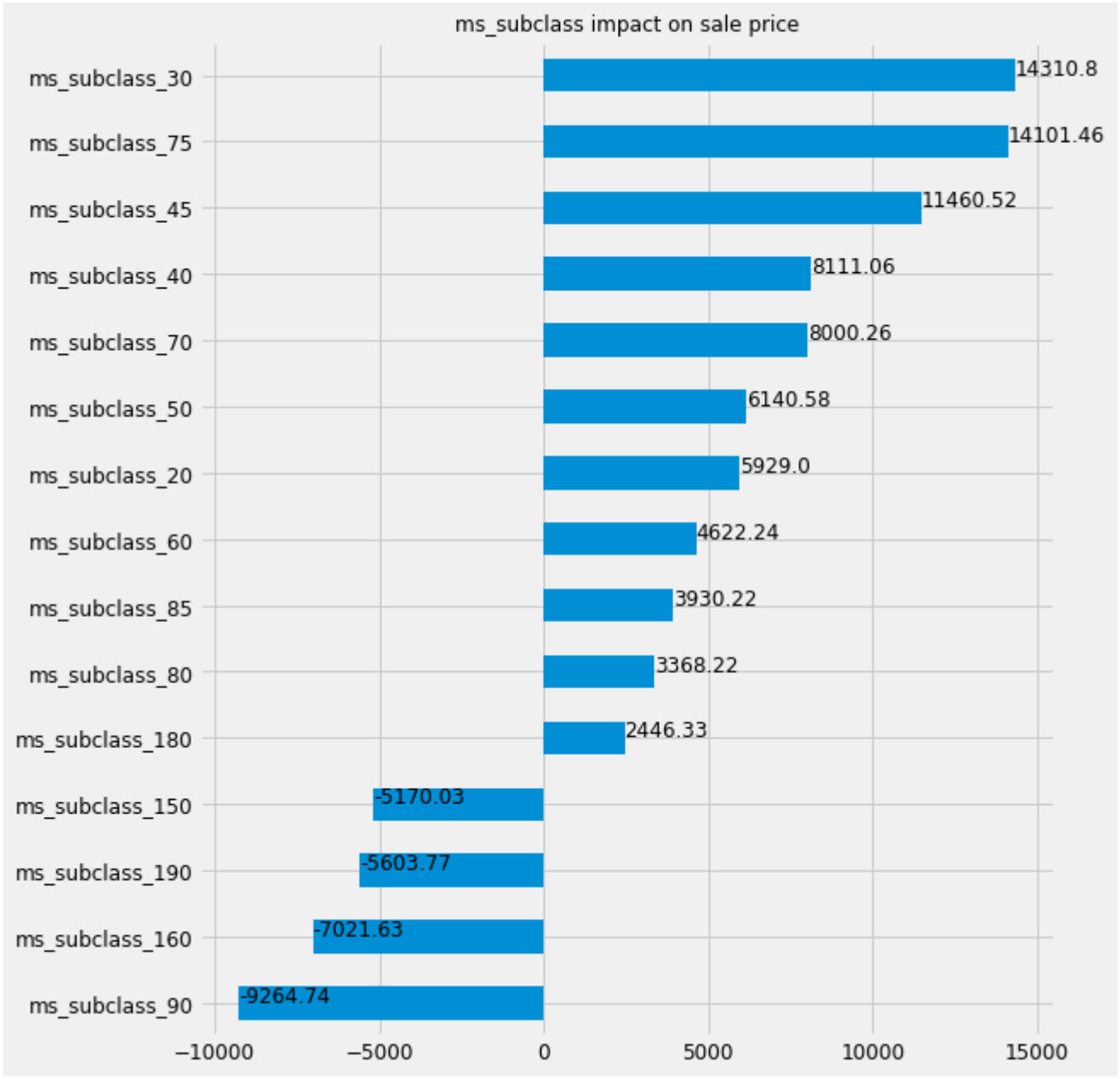
Cross validation (Prediction on unseen data)
R2 Score: 0.9099
Root Mean Square Error: 23734.1034

Interpretation:

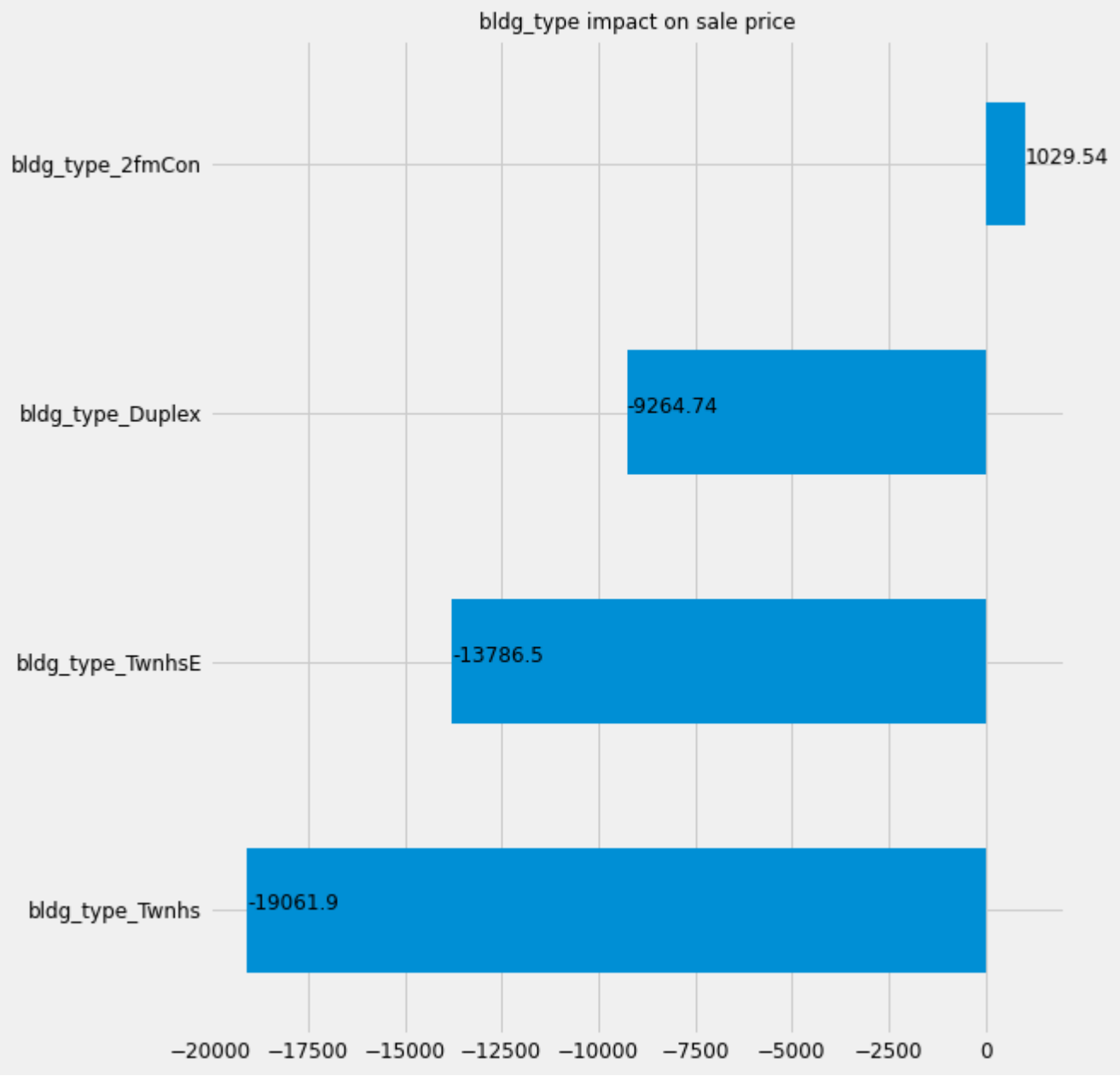
From the above metrics, we can see that model perform better on training data than on unseen data which can be interpreted that model is **slightly overfit or having high variance**. The predict price can have the error interval of +/- 23734 USD which indicates the **low bias** of this model



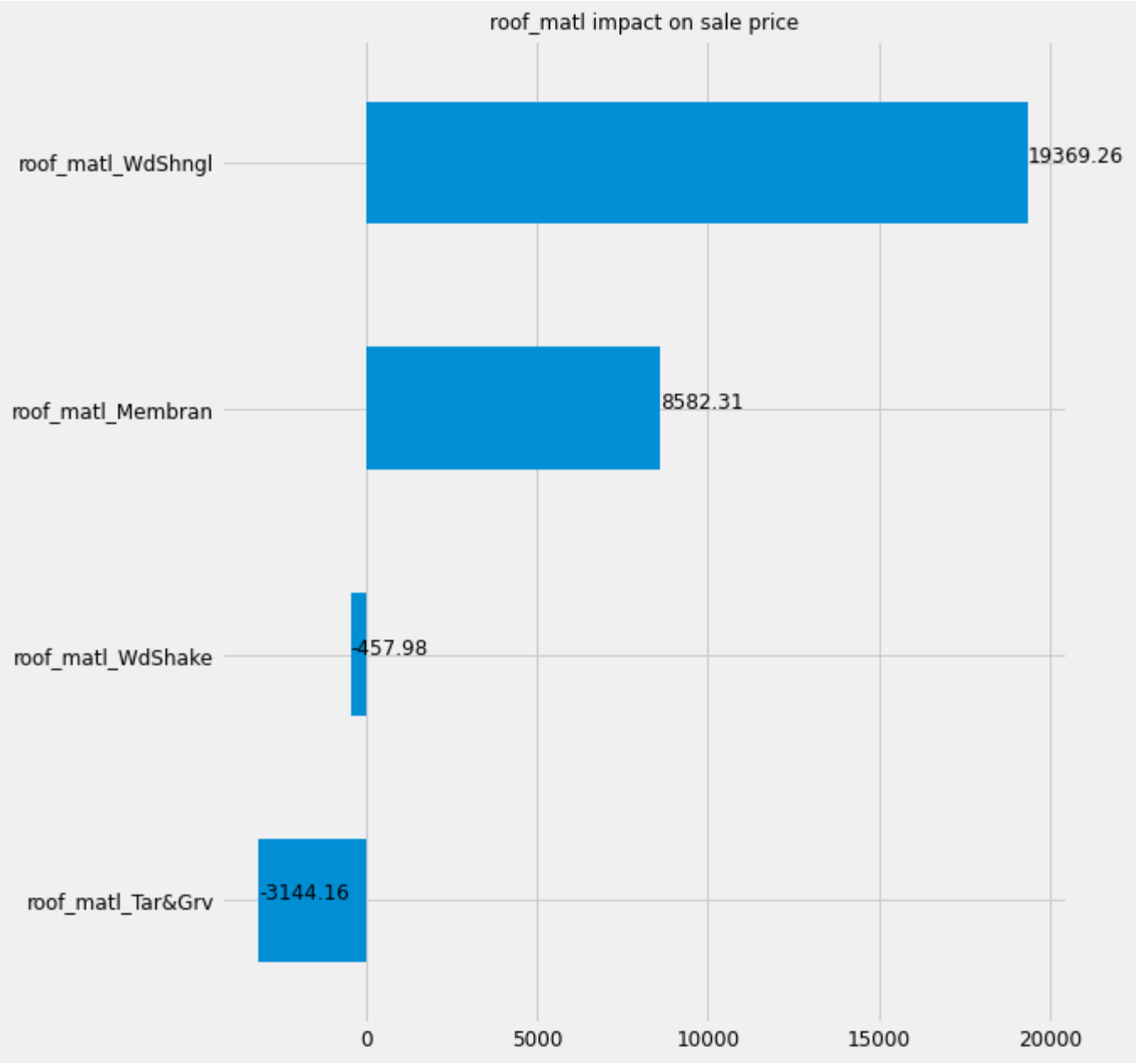
Neighborhood



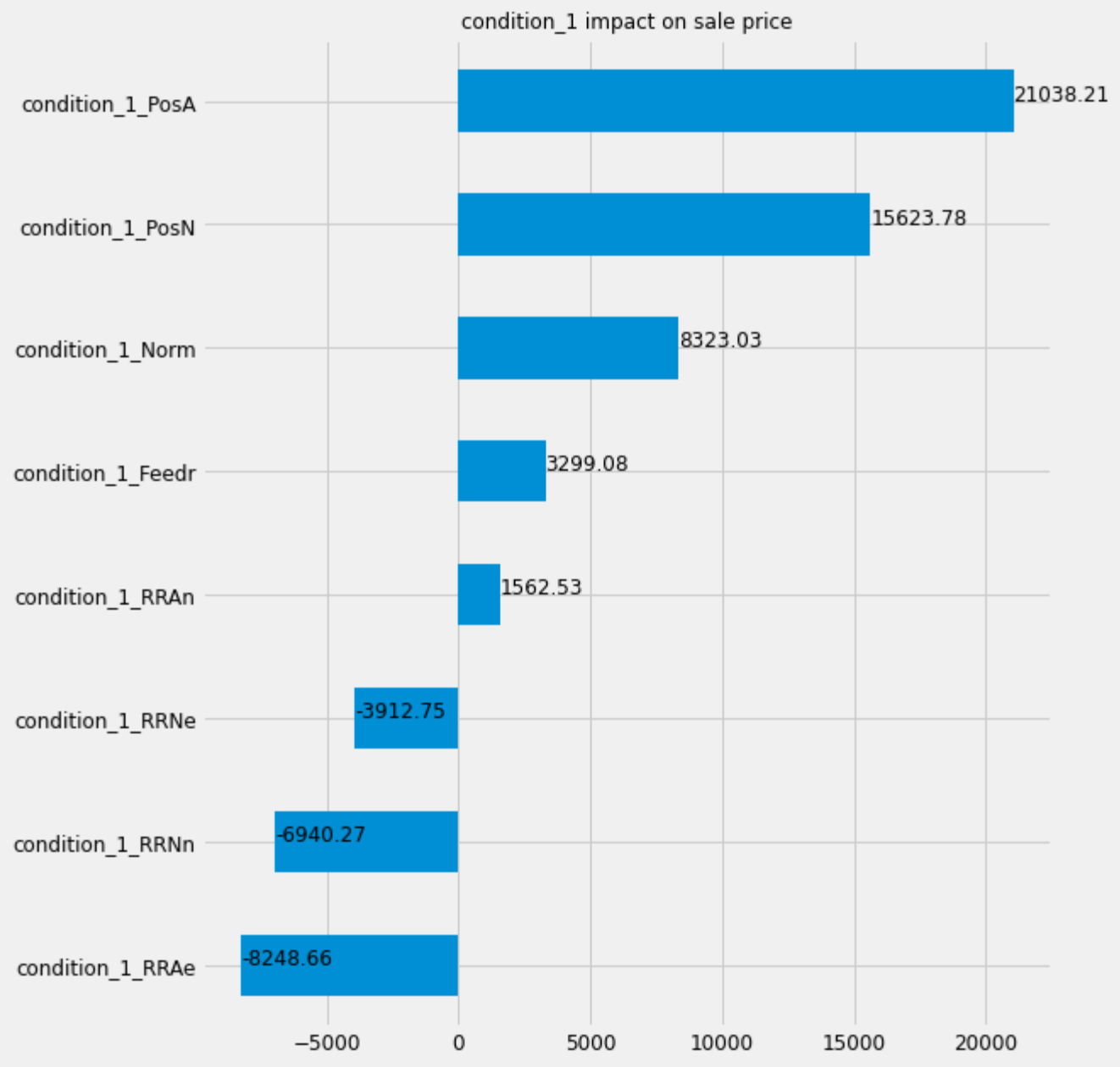
Building Class



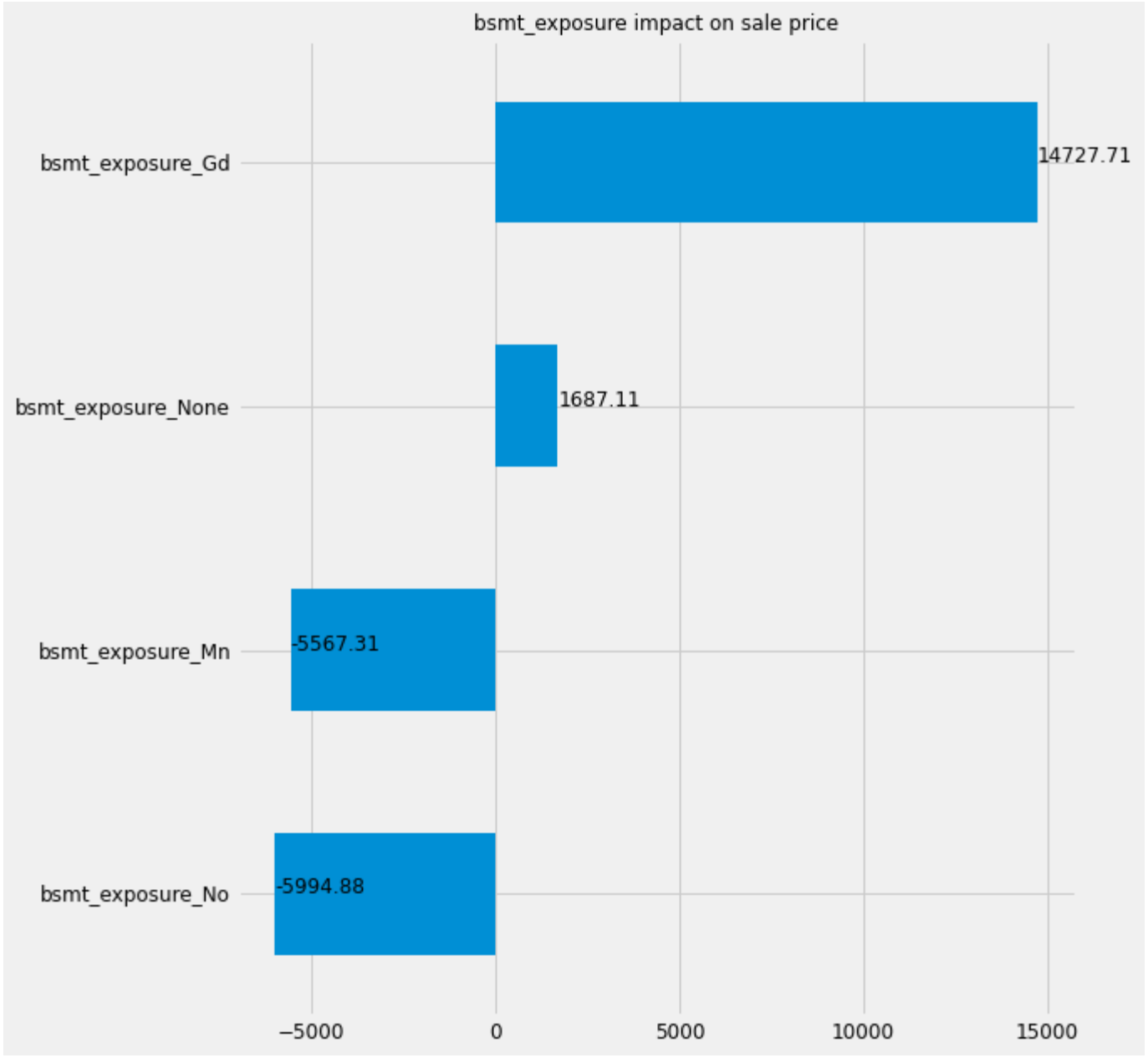
Building Type



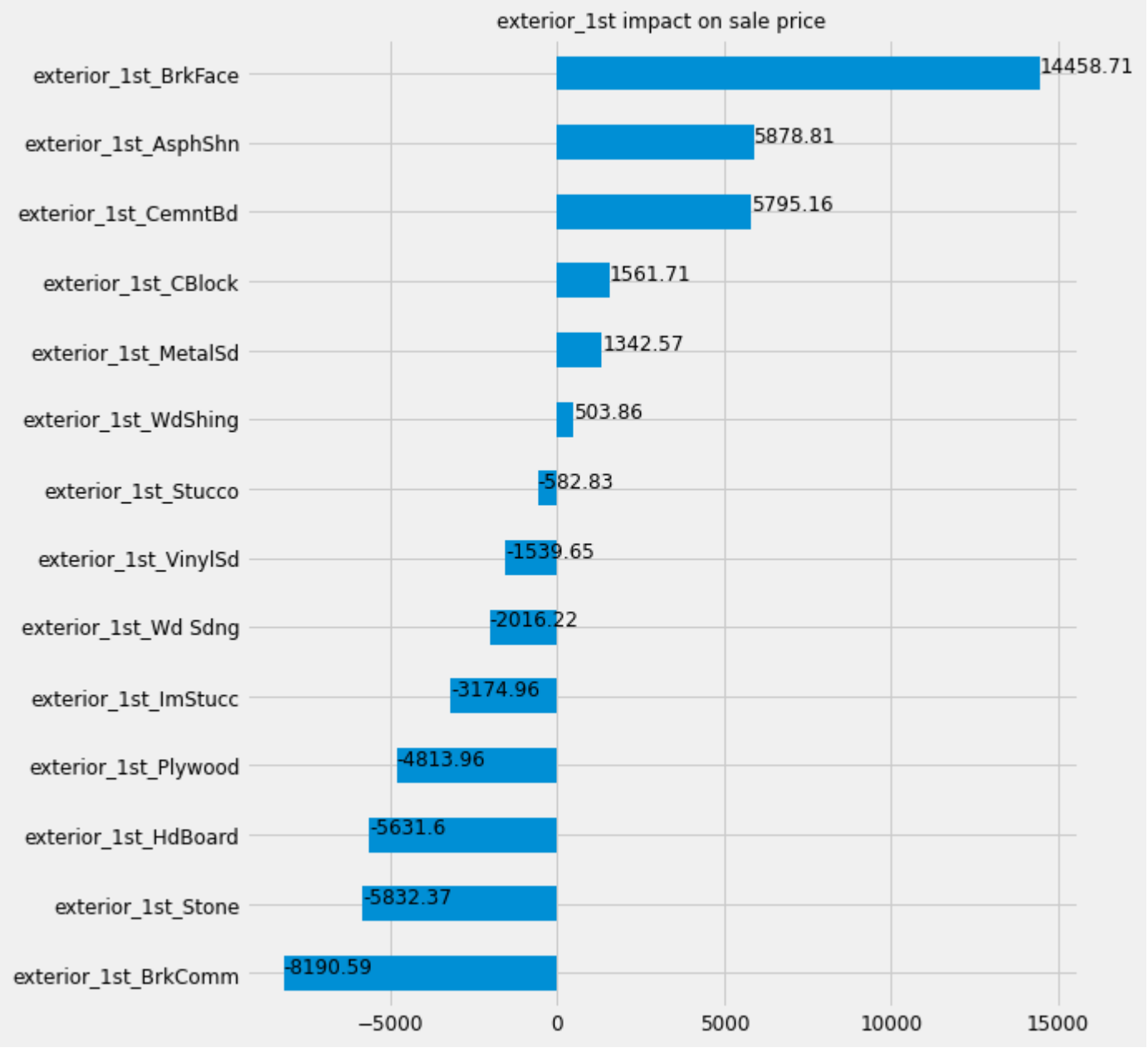
Roof Material



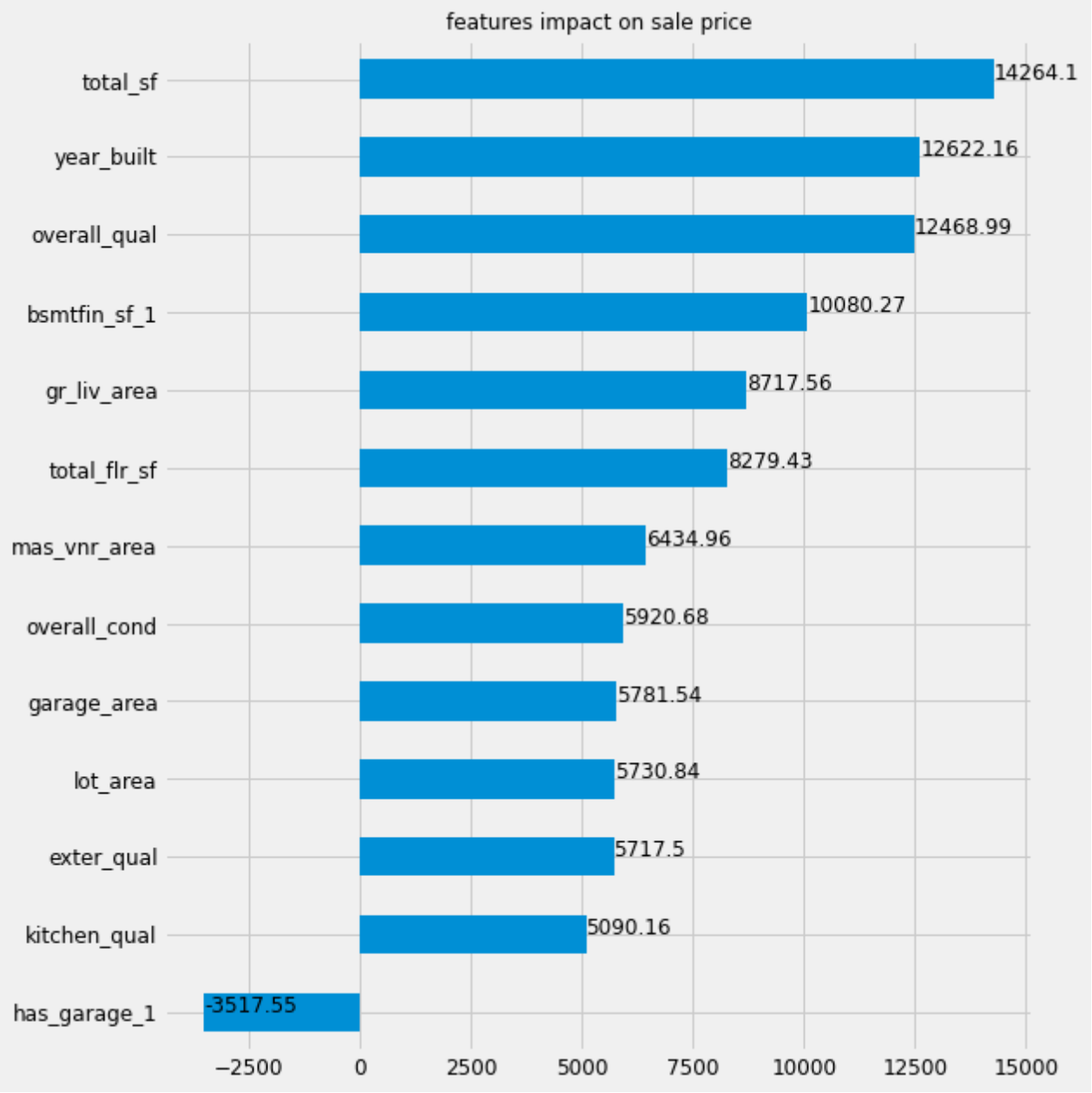
Proximity to main road or railroad



Basement Exposure



Exterior covering on house



Area/Quality of the house

#	Features	High positive impact	High Negative Impact
1	Neighborhood	**Green Hills, Stone Brook and Northridge Heights**	Sawyer West,Timberland, Edwards and College Creek
2	MS Zoning	Floating Village Residential	Commercial Zoning
3	MS Subclass	1-story/1945&older , 2-1/2 story/all ages	duplex-all styles and ages, 2-story pud- 1946&newer
4	Exterior covering on house	**Brick Face**	Brick Common
5	Masonry veneer type	Stone	Brick Face
6	Roof Style	Hip	**Mansard**
7	Roof material	**Wood Shingles**	Gravel & Tar
8	Bldg Type	Two-family Conversion	**Duplex**
9	Heating	Wall Furnace	Hot water or steam heat other than gas
10	Basement Exposure	**Good Exposure**	No Exposure
11	Rating of basement finished area	**No Basement**	Average Rec Room
12	Garage Finish	**No garage**	Rough Finished
13	Home Functionality	Typical Functionality	Severely damaged house
14	Flatness of the property	Hillside	Depression
15	Lot configuration	Cul-de-sac	Frontage on 3 sides of property
16	Proximity to various conditions	Adjacent to postive off-site feature	Adjacent to East-West Railroad

#	Features	Positive/Negative Impact
1	Total Area in Sq.Ft.	Positive
2	Year Built	Positive
3	Overall Qualty	Positive
4	Basement finished area in Sq.Ft.	Positive
5	Above grade (ground) living area in Sq.Ft.	Positive
6	Total Floor Area in Sq.Ft.	Positive
7	Masonry veneer area in Sq.Ft.	Positive
8	Overall Condition	Positive
9	Garage Area in Sq.Ft.	Positive
10	Lot Area in Sq.Ft.	Positive
11	Exterior Quality	Positive
12	Kitchen Quality	Positive
13	has garage	Negative

Conclusion

Based on our problem statement, we found that
1. **Neighborhood and the location** of the house is really matter. If sellers doesn't have the house in particular area, it is hard to rise the price above others house.
2. Using the **right material** and the **right style** can impact your housing price. Wood Shingles as your roof material and if your exterior covering is Brick Face can highly increase the price sold.
3. Make sure that house can **function properly** that basement has good exposure, or electricity is good. If not, the price can be a lot lower.
4. **You don't need to build garage** if you didn't have one. Surprisingly having garage in Iowa can decrease the price!.

Conclusion

Limitation of our prediction

1. The dataset used for train contains only about 2000 data points where the sale price only cover from 12789 USD to 611657 USD. Model will perform badly if the expecting price is out of range.
2. The dataset only contains housing price data in IOWA. If the model is going to be used in other states on country, it can perform badly as well.
3. Now, the model is slightly overfit and the predicted price doesn't not represent the correct price of the house. It can be lower or higher. please use the model wisely.

THANK YOU